A SNP Resource for Human Chromosome 22: Extracting Dense Clusters of SNPs From the Genomic Sequence

Elisabeth Dawson,[1,7] Yuan Chen,[1,7] Sarah Hunt,[1,7] Luc J. Smink,[1] Adrienne Hunt,[1] Kate Rice,[1] Simon Livingston,[1] Suzannah Bumpstead,[1] Richard Bruskiewich,[1] Pak Sham,[2] Rocky Ganske,[3] Mark Adams,[4] Kazuhiko Kawasaki,[5] Nobuyoshi Shimizu,[5] Shinsei Minoshima,[5] Bruce Roe,[6] David Bentley,[1] and Ian Dunham[1,8]

[1]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [2]Section of Genetic Epidemiology and Biostatistics, Department of Psychiatry, Institute of Psychiatry, London SE5 8AF, UK; [3]Third Wave Technologies, Inc., Madison, Wisconsin 53719-1256, USA; [4]The Institute for Genomic Research, Rockville, Maryland 20850, USA; [5]Department of Molecular Biology, Keio University School of Medicine, Shinjuku-ku, Tokyo 160–8582, Japan; [6]Department of Chemistry and Biochemistry, The University of Oklahoma, Norman, Oklahoma 73019, USA

The recent publication of the complete sequence of human chromosome 22 provides a platform from which to investigate genomic sequence variation. We report the identification and characterization of 12,267 potential variants (SNPs and other small insertions/deletions) of human chromosome 22, discovered in the overlaps of 460 clones used for the chromosome sequencing. We found, on average, 1 potential variant every 1.07 kb and approximately 18% of the potential variants involve insertions/deletions. The SNPs have been positioned both relative to each other, and to genes, predicted genes, repeat sequences, other genetic markers, and the 2730 SNPs previously identified on the chromosome. A subset of the SNPs were verified experimentally using either PCR–RFLP or genomic Invader assays. These experiments confirmed 92% of the potential variants in a panel of 92 individuals. [Details of the SNPs and RFLP assays can be found at http://www.sanger.ac.uk and in dbSNP.]

One of the aims of the Human Genome Project is to provide a reference DNA sequence from which to detail the sequence variation that exists in the human population. This genotypic diversity is presumed to underlie the heritable phenotypic differences observed as variation in drug response, susceptibility to disease, and other complex traits. It is hoped that cataloging DNA sequence variation will provide the tools to relate genotypes with complex phenotypes and hence enable discovery of the causative genetic factors (Risch and Merikangas 1996; Collins et al. 1997; Housman and Ledley 1998).

Common types of sequence variation in humans include single nucleotide polymorphisms (SNPs), insertions and deletions of a few nucleotides, and variation in the repeat number of a motif (mini- and microsatellites). SNPs have a very high abundance in the genome, occurring at a density of ~1 per kb when any two genomes are compared, and have a low mutation rate. These characteristics make SNPs potentially valuable markers for association-based approaches to discovering the genetic components of complex traits. Furthermore, where large sample sizes are required, their biallelic nature is amenable to high throughput automated genotyping. In addition, SNPs are potentially useful genetic markers for family-based linkage studies of Mendelian diseases (Kruglyak 1997), studying population histories and genetics (Chakravarti 1999) and personalization of medicine (Housman and Ledley 1998).

Previous SNP finding efforts have focused on (1) candidate genes for common diseases (Nickerson et al. 1998; Cambien et al. 1999; Cargill et al. 1999; Halushka et al. 1999), (2) genes with expressed sequence tags (Wang et al. 1998; Buetow et al. 1999; Garg et al. 1999; Marth et al. 1999; Picoult-Newberg et al. 1999), or (3) genomic sequence (Kwok et al. 1996; Kawasaki et al. 1997; Horton et al. 1998; Lai et al. 1998; Taillon-Miller et al. 1998, 1999; Wang et al. 1998). In the gene-based approaches, variations have been found by either resequencing the gene of interest in a number of individuals or electronically aligning EST/

[7]These authors contributed equally to this work.
[8]Corresponding author.
E-MAIL id1@sanger.ac.uk; FAX 01223-494919.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.156901.

11:170–178 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

cDNA sequences. To find variations in genomic sequence the two predominant methods have been either to resequence the area of interest using *Alu*-PCR (Lai 1998), STS fragments (Kwok et al. 1996; Wang et al. 1998), a reduced representation shotgun library or a sheared genomic library (Altshuler et al. 2000; Mullikin et al. 2000), or to use the overlapping clones from large-scale sequencing projects (Kawasaki et al. 1997; Horton et al. 1998; Taillon-Miller et al. 1998). Although SNPs in coding sequences (cSNPs) are potentially functional and therefore, invaluable in direct, candidate gene-based association studies, our knowledge of the complete set of human genes is still partial and cSNPs are labor intensive to obtain. Alternatively, although genomic SNPs may not lie within genes, they still may be used effectively in indirect association studies that rely on neutral markers being in linkage disequilibrium with the pathogenic variant. Futhermore, progress in human genome sequencing now allows very large numbers of genomic SNPs to be obtained rapidly and economically.

Here we use the overlapping sequence from neighboring clones used for the genomic sequencing of human chromosome 22 (Dunham et al. 1999b) to develop a collection of more than 12,000 candidate variations for this chromosome. The variations are distributed across the chromosome in clusters with a mean density of 1 per 1.07 kb. 82% of the candidate variations are 1 bp substitutions, whereas the remaining 18% involve insertions/deletions (indels). Experimental verification indicated that 92% of the candidate variations (including both SNPs and indels) are present in a panel of 92 individuals. Each variation is mapped to single-base accuracy in the reference sequence of the chromosome, and is placed relative to each other and relative to the annotated genes, predicted genes, and repeat sequences established by computational and biochemical analysis (Dunham et al. 1999b). In addition, because the analysis gives the full sequence flanking each variant that is essential for development of an assay, the SNPs are "genotype ready". In combination with the collection of 2730 randomly distributed SNPs from chromosome 22 described in Mullikin et al (2000), chromosome 22 is now covered with 15,000 SNPs and other small variations.

RESULTS

Extraction of Variation Data

Comparison of overlapping sequence from adjacent genomic clones provides an efficient and economical source of sequence polymorphisms (Taillon-Miller et al. 1998). To identify candidate variations on chromosome 22 we exploited the presence of the different haplotypes represented in the clones used to determine the sequence (bacterial artificial chromosomes [BACs], P1-derived artificial chromosomes [PACs], cosmids, and fosmids). Nine different libraries made from different DNAs were used, representing 17 different haplotypes.

The sequence of chromosome 22 (33.46 Mb) was compiled from the shotgun sequence of 514 overlapping clones. In total, 460 (potentially polymorphic) overlaps representing 15.24 Mb of overlap sequence were investigated (see Methods); 358 overlaps, representing 13.16 Mb of overlap sequence, contained potential variants (SNPs and other small insertions/deletions). Fifty-eight percent of these overlaps involved clones from different libraries (and therefore, represented two different individuals). The remaining overlaps involved clones from within a library representing the two different copies of chromosome 22 from one individual. There were 12,267 candidate variants, of which 10,051 (82.1%) were single base substitutions (Table 1). One thousand three hundred thirty-four (10.9%) were single base insertions/deletions (indels) and the remainder were substitutions or indels extending >1 bp.

The sizes of the overlaps containing variations ranged from 356 to 69,693 bp, with an average of 36,752 bp. For overlaps >10 kb, with potential variants

Table 1. Polymorphism Discovery

		% of total number of candidate SNPs
Number of overlaps looked at	460	
Size of overlaps (bp)	15,241,466	
Number of candidates detected:	12,267	
Number of single bp substitutions	10,051	81.94%
Number of single bp insertions/deletions	1,334	10.87%
Number of insertions/deletions >1 bp	846	6.90%
Number of single bp sub/indel[1]	19	0.15%
Number of sub/indel >1 bp[2]	17	0.14%

Categories of potential variations found in the overlap sequences. Some variations involving adjacent nucleotides are complex and it is difficult to say which nucleotide has been inserted, deleted, or substituted. Therefore, we grouped some variants together as shown.

[1]These are situations where the variation involves 2 bp in one sequence and 1 bp in another—one nucleotide appears to be substituted and the other inserted/deleted, for example, CAT–CTCT.

[2]These are situations where the variation involves >2 bp in one sequence and a different number of base pairs in the other—the different nucleotides appear to be either substituted or inserted/deleted, for example, CATGTGT–CCCT.

(289 overlaps), the average number of variants per kb is 0.93 (1 variant per 1.07 kb); with a range of 0.01 (1 variant per 100 kb) to 5.04 (1 variant per 0.198 kb). Examination of the verification data (see below) for candidate variations in the overlaps with very low densities showed that no candidate variation was confirmed in an overlap with a variant density of <0.07 (1 variant per 15 kb). Therefore, it is unlikely that many of the candidate variations in the overlaps with very low densities are real. Given that the error rate in finished sequence is ~1 in 50,000 bp (Dunham et al. 1999b), and that we are comparing finished sequence to unfinished shotgun sequence, which has a higher error rate (see Methods), this is not unexpected. The overlaps with variant density of <0.07 (1 variant per 15 kb) represent one empirical assessment of the likely level of false positives from cloning and sequencing artifacts, and alignment problems during candidate variation extraction. Taking the verification data into account suggests the range of densities for potential variants is 0.07 (1 variant per 15 kb) to 5.04 (1 variant per 0.198 kb). Other studies, based on the analysis of random genomic sequence, have shown similar densities (Kwok et al. 1996; Kawasaki et al. 1997; Horton et al. 1998; Taillon-Miller et al. 1998, 1999; Wang et al. 1998).

One hundred two overlaps (2.08 Mb) contained no variants; in most cases (84.3%) this was because the two overlapping clones were derived from the same library and therefore, probably from the same haplotype. However, 16 (15.7%) of the overlaps without variants involved overlapping clones from different libraries. These overlaps had a mean size of 36,595 kb consistent with the characteristics of all overlaps. Looking only at overlaps involving clones from different libraries, 6.3% by sequence length had no candidate variation, implying that there is appreciable haplotype sharing in the small sample of individuals from which the different libraries were made. This haplotype identity can extend up to 137 kb. The identities of the individuals used for the libraries involved are not known, but there is no reason to believe that they are close relatives. It would be useful to sequence more large regions from more individuals to determine whether this degree of haplotype identity is typical.

Verification

To assess the usefulness of the collection we used a PCR–RFLP approach to verify a subset of the potential polymorphisms, and to develop a set of biallelic markers. This method is relatively inexpensive, simple, and amenable to any research group who might want to use the biallelic markers. Of the potential variations, 42.5% affected a restriction enzyme site, and 19.3% (of the total) could be converted into a unique PCR-based assay within the confines of the size restrictions we used.

Candidate variants (both SNPs and indels) were randomly selected and screened for suitability for PCR–RFLP assay as described in the Methods. Verification was performed using DNA from a panel of 20 unrelated Caucasians, collected as controls for disease association studies. Of 415 candidate variants that were tested in this manner by PCR–RFLP, 90% were polymorphic in the DNA panel used, and 59% had a lower allele frequency of ≥0.2; a frequency suggested to be valuable for disease gene mapping (Kruglyak 1997). Ten percent did not appear to be polymorphic in the 20 Caucasians we studied. This could reflect low frequency alleles, artifacts from the sequencing/cloning process, or false-positive potential variants. We believe that cloning artifacts must occur at a low frequency compared to the rate of true polymorphism because we observed 2.08 Mb of overlap sequence (102 overlaps) without detectable variation. However, our failure to confirm any potential variations in the overlaps with variant densities of <0.07 (1 variant per 15 kb; see above) suggests that this is the level at which cloning/sequencing artifacts or misalignments during the extraction process become significant.

With 20 individuals, an allele must be present at a frequency of at least 0.025 in our sample to be observed. However, the probability of observing an allele with a general population frequency of 0.025, in a sample of 20 individuals, is only 63%. One of the limitations of this study is that only two chromosomes are being used to find polymorphisms at any point. Hence, if the sequence is derived from a rare haplotype, the alleles may not be common in the population used for verification. Although we do not have access to the DNA of the individuals used to construct the libraries, the existence of some of these low frequency alleles can be tested in larger samples of individuals.

Therefore, 90 additional independent variants were chosen at random and studied in a separate sample of 92 unrelated Caucasian individuals (ECACC, Porton Down, UK) using the Invader assay as the genotyping method (Lyamichev et al. 1999; Ryan et al. 1999; Hall et al. 2000). Eighty-three variants (92%) showed evidence of polymorphism; 22 (24%) of these polymorphic variants had a lower allele frequency of <0.1 (compared to 11% from the 20 individuals) (Table 2), which supports our interpretation that some of the variants that appeared to be nonpolymorphic when tested with RFLP–PCR in the 20 individuals may represent low frequency alleles. Fifty-two percent of the variants tested with the Invader assay alone had a lower allele frequency of between 0.2 and 0.5 (similar to the outcome from PCR–RFLP, 59%). In addition, 31 variants were genotyped with both methods and 30 of 31 variants had concordant results for outcome. One

Table 2. Verification of Polymorphisms

	Lower allele frequency (*P*)	PCR-RFLP[1]	Invader[2]	Total
Number of variants tested		415	90	505
Number of variants with lower allele frequency	$0.4 \leq P \leq 0.5$	77 (19%)	15 (17%)	92 (18%)
	$0.3 \leq P < 0.4$	83 (20%)	23 (25%)	106 (21%)
	$0.2 \leq P < 0.3$	83 (20%)	9 (10%)	92 (18%)
	$0.1 \leq P < -0.2$	83 (20%)	14 (16%)	97 (19%)
	$0.0 \leq P < 0.1$	47 (11%)	22 (24%)	69 (14%)
Number not polymorphic	0	42 (10%)	7 (8%)	49 (10%)

[1]Used 20 individuals for verification.
[2]Used 90 individuals for verification.

variant did not appear to be polymorphic in the 20 individuals with PCR–RFLP, but showed an allele frequency of 0.073 in the larger panel.

In total, we have genotyped 505 independent variants: 285 (57%) of which have a lower allele frequency of between 0.2 and 0.5 (Table 2) and therefore, will be useful for genetic studies. The distribution of allele frequencies appears to be fairly uniform. However, there is a smaller probability for detecting a rare allele than a common allele in a finite sample of individuals; therefore, in reality there may be an excess of rare alleles. This is consistent with a neutral drift model with mutation, as well as observed allele frequency distributions of classic polymorphisms (Falconer and Mackay 1996). The 8% of candidate variations without polymorphism, evident from Invader assays, represents the false-positive rate for the discovery of variants from sequence clone overlaps and/or variants with a very low allele frequency.

Analysis of Base Changes

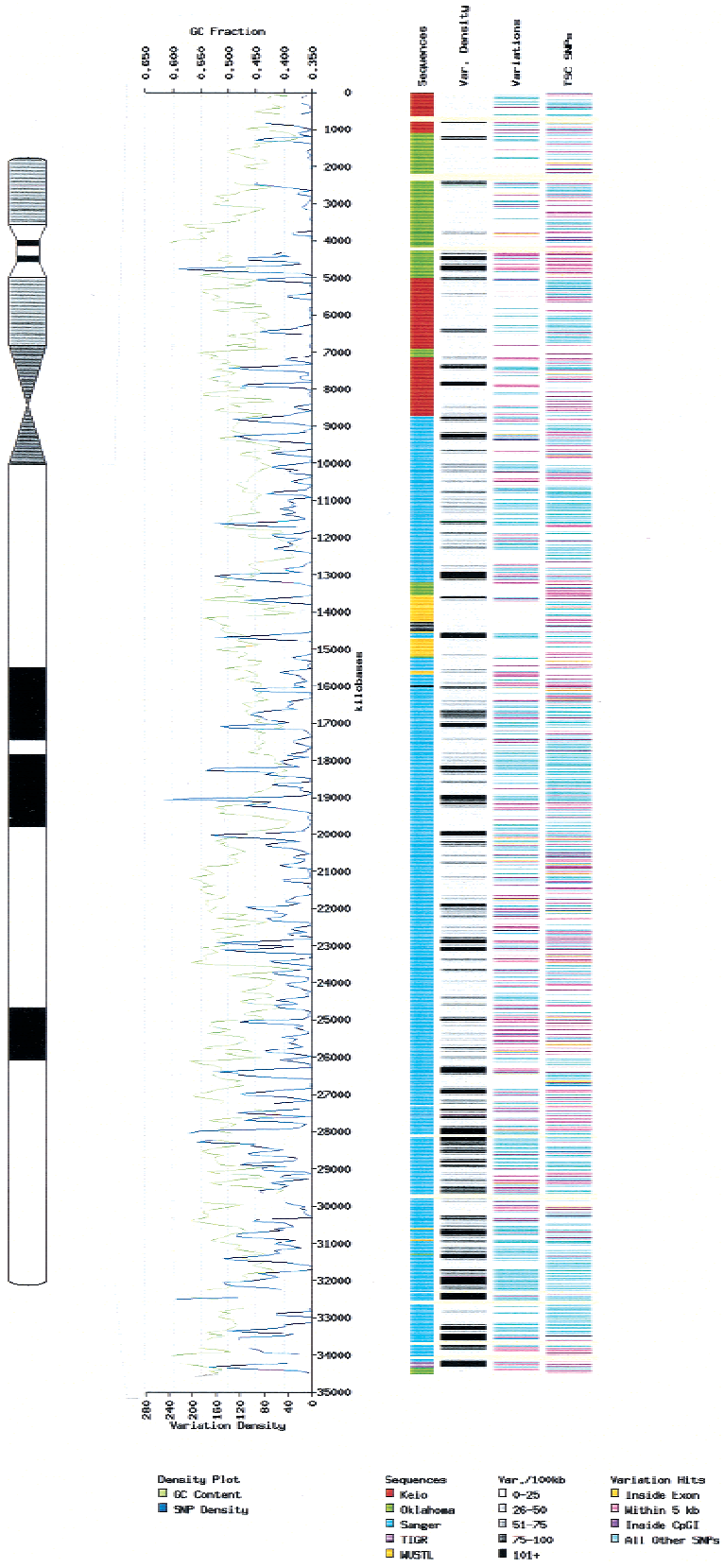
We further investigated the base changes involved in each substitution (Table 3). Under random mutation, half as many transitions as transversions would be expected to occur. We found that transitions occur 2.4 times more often than transversions, a pattern already reported in other SNP identification studies (Kawasaki et al. 1997; Horton et al. 1998; Nickerson et al. 1998; Wang et al. 1998; Cambien et al. 1999; Halushka et al. 1999) and mutation surveys (Vogel and Kopun 1977; Krawczak et al. 1998). The most common change was C/T (or G/A) (70.2%). The higher level of C/T (G/A) SNPs probably reflects the deamination of 5-methylcytosine that occurs frequently at CpG dinucleotides (Holliday and Grigg 1993) and results in converting cytosine to thymine. Analysis of the nucleotide neighbors of the proposed substitutions reveals a bias for C/T (G/A) substitutions to occur at CG dinucleotides [43.4% of C/T (G/A) were at CG dinucleotides]. However, limited conclusions could be drawn because the methylation status of the dinucleotides or the direc-

Table 3. Base Pair Changes

Type of substitution	No. of substitutions
Transitions[1]	
C/T (=T/C, G/A, A/G)	7060 (70.24%)
Total transitions	7060 (70.24%)
Transversions[2]	
C/G (=G/C)	949 (9.44%)
T/A (=A/T)	496 (4.94%)
T/G (=G/T, C/A, A/C)	1546 (15.38%)
Total transversions	2991 (29.76%)

Each base change observed in the candidate variants was categorized to be one of the four types shown.
[1]Transitions are pyrimidine–pyrimidine or purine–purine substitutions.
[2]Transversions are purine–pyrimidine or pyrimidine–purine substitutions.

Figure 1 (following page) Distribution of polymorphisms on human chromosome 22. An ideogram of chromosome 22 with a schematic representation of the Giemsa banding pattern is shown at left. Next, the region containing the finished sequence is expanded to show the SNP map. The SNP density (the number of candidate variants in consecutive 100-kb regions) is plotted (blue line) superimposed on a plot of GC density (green). GC content is calculated as a percentage of the sequence using a sliding 100-kb window moved in 50-kb increments. The first column to the right of the graph represents sequences color-coded as per the collaborating institutions that contributed to the sequence pale yellow bars drawn horizontally across the map represent current gaps in the completed sequence. The next column to the right (Var. Density) shows a gray scale coding the number of potential variations recorded in the given 100-kb region. Such potential variations may be SNP, insertion, or deletion polymorphisms relative to the published reference sequence. The next column represents these variations as line annotations with color coding to represent the position of the variation relative to genomic features such as exons and CpG islands (a color key is on the diagram). The last column represents the corresponding map of recently published TSC (The SNP Consortium) SNPs reported by our group. The very high density of the overlap variations described in this paper (the Variations column) is not evident from the diagram because of the limits of the resolution. This diagram can also be viewed as a link from http://www.sanger.ac.uk/cgi-bin/humace/snp_search where a zoom facility allows regions to be enlarged to show the positions of individual SNPs relative to annotated exons and CpG islands.



tion of the primary mutation event is unknown. We hypothesized that, just as with the known spectrum of sequence variants that underlie disease (Krawczak et al. 1998), the abundance of C/T (G/A) SNPs will perhaps be higher in the gene and C + G-rich isochores. The relative frequencies of the four SNP types were correlated with GC content, and we found that G/C SNPs were relatively more abundant in the GC-rich regions [odds ratio (OR) = 1.11 per 5% increase in GC content, 95% confidence interval (CI) = 1.06–1.15, $P < 0.00001$] and A/T SNPs less so (OR = 0.83, 95% CI = 0.79–0.88, $P < 0.00001$), as would be expected given the base composition. The number of C/T SNPs was not significantly positively correlated with high GC. This may be because the SNP discovery is derived, at each point, from only two haplotypes and the observed SNP-type density may not be representative of the full spectra in human populations. Given the increased level of C/T changes it would be expected that the remaining transversions occur at a similar level to each other. However, it should also be noted that A/T changes are underrepresented.

Positions of Potential Variants

The clustered distribution of candidate variants along the length of the chromosome (Fig. 1) reflects the method of discovery. As expected, 10,584 (86.3%) of the candidate variants are <2 kb away from another candidate variant. Large clusters of potential variants represent regions where there are nearly full clone overlaps because the region has been sequenced twice, by different sequencing centers. There are 78 gaps between clusters of candidate variants >100 kb. Ten of these are the gaps in the chromosome sequence. The others represent regions where no variants could be extracted because the overlapping clones have identical haplotypes [e.g., the immunoglobulin-like (IGL) region (5000- to 7000-kb region on Fig. 1) was predominantly sequenced using only a single haplotype cosmid library] or there was little available overlap sequence (e.g., the sequences from Washington University, 13,500- to 15,300-kb region on Fig. 1). The largest distance between candidate variants is 88,3716 bp.

We observed that 221 potential substi-

tutions were adjacent to another potential substitution, forming 101 potential double nucleotide substitutions (and 5 potential triple and 1 potential quadruple nucleotides substitutions). Under random occurrence of SNPs, the probability that a nucleotide next to a SNP is also a SNP is 0.0008; therefore, the expected numbers of single, double, triple, and quadruple SNPs are 12217.36, 9.81, 0.01 and 0.00, respectively, which are significantly different from the observed values (χ^2 965.4, df = 2). The observed rate of multiple nucleotide substitutions is consistent with a probability of between 0.007 and 0.011 for a nucleotide next to a SNP to be also a SNP, representing ~10-fold increase from the overall rate of SNP occurrence. An increased level of double-nucleotide substitutions has also been observed in other primate noncoding sequences (Averof et al. 2000). There are 41 possible double-nucleotide substitutions, but we observed only 31 of these. The TG ⇔ CA type was the most common (n = 19) and more than twice as frequent as any other type. Because we chose candidate variations for verification randomly, none of the pairs of substitutions has yet been tested, and this awaits further study. The number of double substitutions that occur within repeats (43%) is similar to that for all candidate variants (see below), suggesting that there is no bias toward or against these substitutions occurring in repeat sequences.

The candidate polymorphisms found on human chromosome 22 represent variation in genomic sequence. The complete sequence of this human chromosome consists of 41.6% repeat sequences and 3% gene (exon) sequence. Here, 47.2% (5788) of the candidate variants occur in repeat sequences [3128 (25.5%) are in Alu] and 1.6% (192) occur in exons. The relative number of candidate variants in exons (1.6% of total) compared to the proportion of the chromosome that contains genes (exon sequence) is consistent with greater selection constraint on genic regions.

In addition, 4303 (35%) candidate variants are within 5 kb of an exon, and 7985(65%) are within 25 kb of an exon, thus may be of use for linkage disequilibrium studies (assuming that linkage disequilibrium extends this far in these regions). Three hundred seventy-two (68%) of the 545 annotated genes have a SNP within 25 kb.

DISCUSSION

As a byproduct of the Human Genome Project sequencing, a large number of genomic variants of chromosome 22 have been discovered. We have looked at 15.24 Mb of overlap sequence and found 12,267 candidate variants. There was no requirement for new sequencing to be done making this an economical method for polymorphism discovery.

In the present study, the tiling path of clones was not chosen specifically for the purpose of polymorphism discovery and there are 78 gaps between candidate variants that are >100 kb. Although 10 of these gaps are the known gaps in the sequence, the others represent regions of overlapping clones with an identical haplotype or regions where minimal overlap sequence was available. To optimize polymorphism discovery in the future it would be beneficial if genomic sequencing tile paths could be chosen so that overlaps always contained different haplotypes.

For polymorphisms to be useful in human genetic studies they must be assembled into maps. By using genomic sequence for polymorphism discovery the variants found are already mapped, both relative to each other and to the genes within the sequence. One of the main uses of biallelic markers in human genetics will be to discover the genetic components of complex disease using neutral variants for linkage (Kruglyak 1997) or linkage disequilibrium (LD) (Risch and Merikangas 1996). In this present study, by using the clone overlaps as a source of variants, a degree of regularity to the distribution along the chromosome that should be sufficient for linkage studies of families, has been obtained. However, the separation required for genomewide LD studies has not yet been tested empirically for large genomic regions. Simulation studies (Kruglyak 1999) have predicted that a density of 1 SNP per 3 kb may be required. The clusters of high-density SNPs described here, in combination with the less dense but more randomly distributed SNPs from The SNP Consortium (Mullikin et al. 2000) results in a resource that allows the empirical testing of such predicted densities. The variants are being used in studies of chromosome 22 to determine the extent of linkage disequilibrium and how it is affected by marker allele frequency, population size, population history, and genomic environment.

Although the idea of genomewide linkage disequilibrium studies using SNPs is appealing, the reality is that such studies are not yet feasible. In addition to requiring a high SNP density, the other difficulties include the lack of total genomic coverage of useful SNPs, the sample sizes needed, the interpretation of the results, and the economics of such an approach. Currently, random SNPs, such as the resource we have described are useful genetic markers for fine mapping in positional cloning projects. The variants found on chromosome 22 are of sufficient density and validity to be useful for such genetic studies and are currently being used for traits mapped to the chromosome such as schizophrenia (Schwab and Wildenauer 1999) and cognitive ability (Hill et al. 1999).

Most genotype methods rely on PCR amplification to provide enough target for a signal relating to each allele. The need for a PCR step means that primers have to be designed that allow amplification of the DNA

containing the variant. A further advantage in using the genomic sequence as a resource for finding polymorphisms means that there is little limitation to the sequence from which primers can be designed. However, the abundance of repetitive DNA can affect the ability to generate primers to amplify a given variant. For instance, we were not able to obtain PCR–RFLP assays for 25.4% of the potential variants that create/disrupt a restriction site because repetitive DNA prevented primer design within our size constraints. Alternatively, genotyping methods that do not require a PCR step do exist (Baner et al. 1998; Lyamichev et al. 1999), and we were able to test the validity of 90 potential variants using such a method. All of the data relating to the primers/enzymes/oligonucleotides used for the validation experiments in this study can be found in the WWW versions of the chromosome 22 database (http://www.sanger.ac.uk/HGP/CHR_22 or http://www.hgmp.mrc.ac.uk). A search tool is available for information on the positions of SNPs relative to the known genes, markers, other SNPs, and sequence clones (http://www.sanger.ac.uk/cgi-bin/humace/snp_search). In addition, the STS and polymorphism data has been deposited in dbSTS (http://www.ncbi.nlm.nih.gov/dbSTS/) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) where further protocol details can be found.

Exploitation of clone overlaps in the rest of the genome sequence should yield a rich source of variants. Although the use of overlaps from a sequence tile path limits the number of chromosomes sampled and the physical positions of the variants observed, the SNP collection developed in this way will be augmented by other complementary approaches. These include large-scale random methods of SNP detection such as the SNP Consortium (Marshall 1999), the US National Institutes of Health Funded program (Marshall 1997), and whole genome shotgun sequence data (Venter et al. 1998). The benefit of the combination of different approaches is illustrated in Figure 1, which shows how the SNPs from TSC on chromosome 22 (Mullikin et al. 2000) provide SNPs in regions where no overlap variants were found. Currently, the number of SNPs from these more random approaches is increasing, thereby improving the utility of the overall SNP map. The main difference between the resource described here and these more random approaches is that this SNP resource is nonrandom and able to provide much higher densities at relatively little cost. Comparison of the candidate variations identified by the overlap-based method and the random approach of the TSC within the regions covered by overlaps showed that 754 of the 2730 TSC SNPs fell into the overlaps, and 366 of these were identified by both the overlap and random methods. The remaining 11,901 overlap variations and 388 TSC SNPs were unique to their respective discovery method. This amply illustrates that the overlap-based resource provides a very high density of variations within the targeted region. On the other hand, because only two haplotypes are sampled, additional variations will continue to be found by the random approach using more haplotypes. Finding SNPs in clusters, in the clone overlaps gives greater choice of SNPs to assay, for a given area. It may also be useful to cluster SNPs for family-based linkage studies, and thereby obtain the same amount of informativeness as seen from microsatellites (McCarthy and Hilfiker 2000). For these reasons we believe that it is imperative to continue to supplement the genomewide random approaches with extraction of variations from the overlaps of clones provided by the working draft genomic sequence.

METHODS

Extraction of Potential Variants

The sequence of human chromosome 22 was determined from a set of minimally overlapping clones chosen from a map of the chromosome that contains clones from nine different libraries representing 17 different haplotypes. (One of the libraries, the Lawrence Livermore cosmid library, contained only one haplotype because it was made from a cell line containing only one copy of chromosome 22.) Each clone was shotgun-sequenced and then finished to a high degree of accuracy (<1 error in 50,000 bases) by resolving any sequence problems found in the assembly of the shotgun sequences of each clone (Dunham et al. 1999b). Typically only 100 bp of both clones in an overlap were finished, although the overlaps between clones vary in size (average, 38 kb). Therefore, to extract the potential variants from the total overlap, we reassembled the sequence of the clone contributing the unfinished section of the overlap, using the assembly program PHRAP (P. Green; http://www.genome.washington.edu/UWGC/analysistools/phrap.htm). Tandem repeats were masked to optimize alignment of both sequences, thereby reducing the number of false-positive variants between the sequences. To avoid repeat regions we only compared sequence when the matching length between unfinished contigs and finished sequence was >2000 bp. The sequences of both versions of the overlaps were aligned using CROSSMATCH (P. Green; http://www.genome.washington.edu/UWGC/analysistools/phrap.htm). A potential variant was identified when a sequence difference was observed between the aligned sequences of the two clones and both base calls were of a high quality (i.e., the unfinished sequence has a phrap quality value Q >30 and the finished sequence is assumed to be Q >40).

Verification of Potential Variants

PCR–RFLP

To develop the assays, SNPs (or small insertions/deletions) were identified that create/disrupt a restriction enzyme site. Then, 400 bp of sequence was extracted from around the candidate variant, avoiding additional similar restriction enzyme sites. Primers were designed from this sequence, after masking high and medium copy repeats (A.F.A. Smit and P. Green; http://ftp.genome.washington.edu/RM/RepeatMasker.html),

using WI PRIMER V0.5 (Mark. J. Daly, Steve E. Lincoln, Eric S. Lander; http://www-genome.wi.mit.edu/ftp/pub/software/primer.0.5,) with size restrictions so that the product and digestion products could be assayed by 2.5% agarose gel electrophoresis (digested product sizes of between 50 and 350 bp). The primer pairs were also checked to ensure that they had no variants within them.

Optimal PCR conditions were obtained using standard reaction mixes (Dunham et al. 1999a) and varying the annealing temperature. PCR reactions (15 µL volumes) containing 50 ng of DNA were performed on a Tetrad (MJ Research) and 5 µL of the product evaluated on a 2.5% agarose gel before RFLP analysis. Amplified products (10 µL) were digested with the appropriate restriction enzyme (New England Biolabs) in a final volume of 20 µL and incubated at the appropriate temperature for at least 1 h. Unmethylated λ DNA (New England Biolabs) was used as a control for restriction enzyme activity. Digested products were visualized by electrophoresis on a 2.5% agarose gel containing ethidium bromide followed by ultraviolet transillumination and photography. Genotypes were scored manually.

Invader

The Invader assays were designed using the InvaderCreator automated probe design software. To facilitate automated probe design, the 50 bp regions surrounding candidate variants were screened to avoid repeated sequences, nonunique flanking sequences, di- and trinucleotide repeats, other SNPs, or a GC content <20% or >80%. Sequences that met these criteria were selected for design and synthesis of oligonucleotides (one invader probe and two allele-specific signal probes with reporter arms). The genotyping was performed in duplicate using 100 ng of genomic DNA per reaction. Typically, the genomic DNA was predenaturated for 5 min at 95°C and added to a 96-well format microtiter plate containing dried-down buffer, Cleavase enzyme, and FRET probe (Third Wave Technologies) in a volume of 5 µL. A 10-µL mix containing 11.5 mM $MgCl_2$, 0.1 mM probe oligonucleotide, and 1 mM allele-specific signal probe oligonucleotide was added to the appropriate wells. Liquid handling was carried out with the Genesis150 robot (Tecan). The reactions were incubated at 65°C for 4 h on Tetrads (MJ Research), the fluorescence determined using a Cytofluor 4000 (Perseptive), and the data analyzed in Microsoft Excel.

Analysis of Base Changes and Distribution of SNPs Along the Chromosome

The details of the potential variations were stored in ACEdb (Durbin and Thierry-Mieg 1991). Custom Perl scripts using AcePerl (Stein, http://stein.cshl.org/aceperl) were used to retrieve the data then perform the various computations. These computations included the extraction of the base change frequencies and the GC content of the 500 bp on either side of the variant. Using Perl modules to manipulate GFF (General Feature Format) data (http://www.sanger.ac.uk/Software/formats/GFF), the positions of the variations were also compared to gene exons, repeats, and other genomic features. The intervariation distances and genomic densities of the variations were also computed. The relative frequency of each type of SNP, relative to the others, was correlated to GC content by logistic regression analysis using SPSS statistical software. A Perl suite of scripts (the "SuperMap" map drawing facility; R. Bruskiewich and T. Hubbard, unpubl.) were used to draw the graphic map representation of these data, as presented in this paper.

Statistical Analysis of SNP Distribution

Let there be probability P that a nucleotide next to a SNP is also a SNP. If SNPs occur randomly then P is simply the overall frequency of SNPs in the region. The probability that a SNP will belong to a string of exactly c consecutive SNPs is $cP^{c-1}(1-P)^2$, becauses c-1 SNPs must occur before the string is terminated by two non-SNPs, and the original SNP can occupy c different positions within the string of SNPs. The expected number of strings of size c, in a region containing n SNPs, is therefore $nP^{c-1}(1-P)^2$. For a χ^2 test, neighboring cluster sizes are combined to ensure that all expected counts exceed 5. A confidence interval for P can be obtained by altering P to find values at which the χ^2 statistic changes from being significant to nonsignificant.

ACKNOWLEDGMENTS

We thank Helga Koch, Bernard Freeman, and Ian Craig from the Institute of Psychiatry in London (IOP) for the supply of 20 Caucasian DNA samples on which to test the SNPs. The project was supported by the Wellcome Trust; the Fund for Human Genome Sequencing Project of Japan Science and Technology (JST) Corporation, and a Fund for the "Research for the Future" Program from the Japan Society for the Promotion of Science (JSPS) (N.S., S.M., K.K.); and the NIH-NHGRI (B.R.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altshuler, D., Pollara, V.J., Cowles, C., Van Etten, W.J., Baldwin, J., Linton, L., and Lander, E.S. 2000. A human SNP map generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Averof, M., Rokas, A., Wolfe, K.H., and Sharp, P.M. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287:** 1283–1286.

Baner, J., Nilsson, M., Mendel-Hartvig, M., and Landegren, U. 1998. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res.* **26:** 5073–5078.

Buetow, K.H., Edmonson, M.N., and Cassidy, A.B. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet.* **21:** 323–325.

Cambien, F., Poirier, O., Nicaud, V., Herrmann, S.M., Mallet, C., Ricard, S., Behague, I., Hallet, V., Blanc, H., Loukaci, V. et al. 1999. Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65:** 183–191.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalayanaraman, N., Nemesh, J., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22:** 231–238.

Chakravarti, A. 1999. Population genetics—Making sense out of sequence. *Nature Genet.* **21:** 56–60.

Collins, F.S., Guyer, M.S., and Charkravarti, A. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278:** 1580–1581.

Dunham, I., Dewar, K., Kim, U.-J., and Ross, M.T. 1999a. Bacterial cloning systems. In *Genome analysis: A laboratory manual series, Volume 3: Cloning systems* (ed. B. Birren, E.D. Green, S. Klapholz, R.M. Myers, H. Riethman, and J. Roskams), pp. 1–86. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Dunham, I., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M.,

Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., et al. 1999b. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.
Durbin, R. and Thierry-Mieg, J. 1991. A *C. elegans* database available from http://www.acedb.org.
Falconer, D.S. and Mackay, T.F.C. 1996. In *Introduction to quantitative genetics*, Fourth edition. Chapter 4, pp. 78–81. Longman, Harlow, England.
Garg, K., Green, P., and Nickerson, D.A. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res*. **9:** 1087–1092.
Hall, J., Eis, P., Law, S., Reynaldo, L., Prudent, J., Marshall, D., Allawi, H., Mast, A., Dahlberg, J. Kwiatkowski, R., et al. 2000. Sensitive detection of DNA polymorphisms by the serial invasive signal amplification reaction. *Proc. Natl. Acad. Sci.* **97:** 8272–8277.
Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22:** 239–247.
Hill, L., Craig, I.W., Asherson, P., Ball, D., Eley, T., Ninomiya, T., Fisher, P.J., Turic, D., McGuffin, P., Owen, M.J., et al. 1999. DNA pooling and dense marker maps: A systematic search for genes for cognitive ability. *Neuroreport* **10:** 843–848.
Holliday, R. and Grigg, G.W. 1993. DNA methylation and mutation. *Mutat. Res*. **285:** 61–67.
Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J., and Beck, S. 1998. Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.* **282:** 71–97.
Housman, D. and Ledley, F.D. 1998. Why pharmacogenomics? Why now? *Nat. Biotechnol*. **16:** 492–493.
Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J.L., Wang, J., and Shimizu, N. 1997. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.* **7:** 250–261.
Krawczak, M., Ball, E.V., and Cooper, D.N. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* **63:** 474–488.
Kruglyak, L. 1997. The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* **17:** 21–24.
———. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22:** 139–144.
Kwok, P.Y., Deng, Q., Zakeri, H., Taylor, S.L., and Nickerson, D.A. 1996. Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics* **31:** 123–126.
Lai, E., Riley, J., Purvis, I., and Roses, A. 1998. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54:** 31–38.
Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W., Takova, T., Kwiatkowski, R.W., Sander, T.J., de Arruda, M., Arco, D.A., et al. 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol.* **17:** 292–296.
Marshall, E. 1997. "Playing chicken" over gene markers. *Science* **278:** 2046–2048.
———. 1999. Drug firms to create public database of genetic mutations. *Science* **284:** 406–407.
Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nature Genet.* **23:** 452–456.
McCarthy, J.J. and Hilfiker, R. 2000. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat. Biotechnol.* **18:** 505–508.
Mullikin, J., Hunt, S., Cole, C., Mortimore, B., Rice, C., Burton, J., Matthews, L., Pavitt, R., Plumb, R., et al. 2000. A SNP map of human chromosome 22. *Nature* **407:** 516–520.
Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E., and Sing, C.F. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19:** 233–240.
Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., and Boyce-Jacino, M. 1999. Mining SNPs from EST databases. *Genome Res*. **9:** 167–174.
Risch, N. and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.
Ryan, D., Nuccie, B., and Arvan, D. 1999. Non-PCR-dependent detection of the factor V Leiden mutation from genomic DNA using a homogeneous invader microtiter plate assay. *Mol. Diagn.* **4:** 135–144.
Schwab, S.G. and Wildenauer, D.B. 1999. Chromosome 22 workshop report. *Am. J. Med. Genet.* **88:** 276–278.
Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., and Kwok, P.Y. 1998. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. *Genome Res*. **8:** 748–754.
Taillon-Miller, P., Piernot, E.E., and Kwok, P.Y. 1999. Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res*. **9:** 499–505.
Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., and Hunkapiller, M. 1998. Shotgun sequencing of the human genome. *Science* **280:** 1540–1542.
Vogel, F. and Kopun, M. 1977. Higher frequencies of transitions among point mutations. *J. Mol. Evol*. **9:** 159–180.
Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077–1082.

Received July 24, 2000; accepted in revised form September 22, 2000.